Question Answering on SQuAD with QANet

Conor McAvity

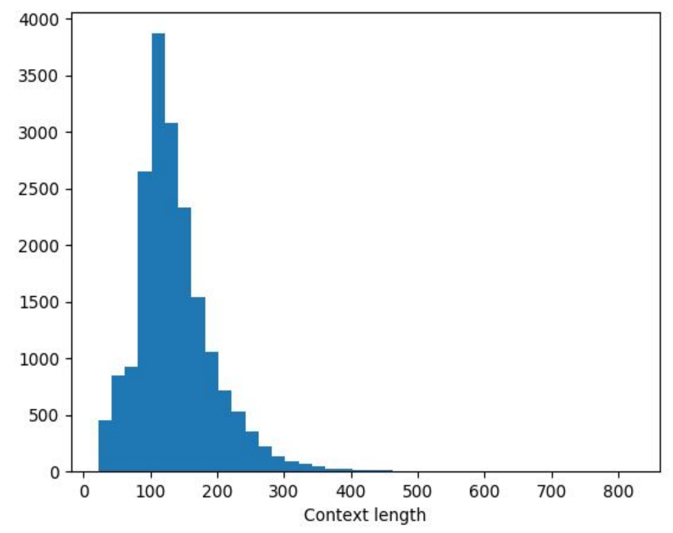
Problem

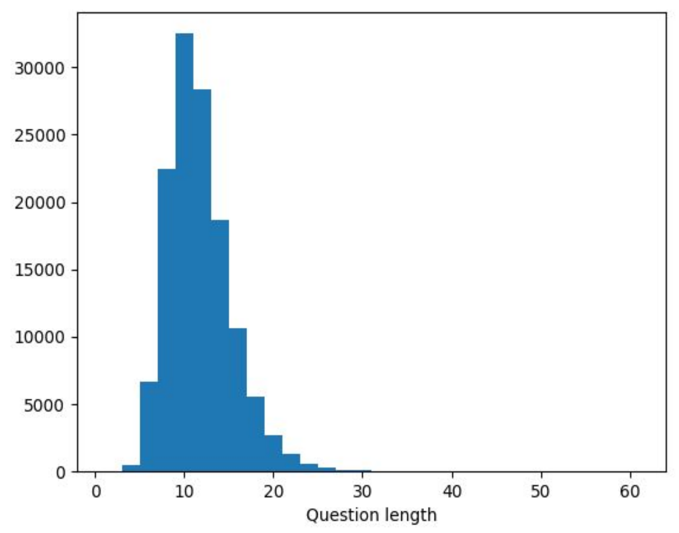
Machine reading comprehension is a task in which a system is asked a question about a passage of text, and needs to find the answer within the text. A successful reading comprehension system has a wide array of real-word applications, such as in digital assistants.

The traditional approach to this task was to build hand-crafted categorical feature classifiers. In the last 5 years, end-to-end neural models have had considerable success.

Data and Task

The dataset used was SQuAD 2.0, the Stanford Question Answering dataset, which consists of 130,000 crowd-sourced (possibly unanswerable) questions about Wikipedia passages.





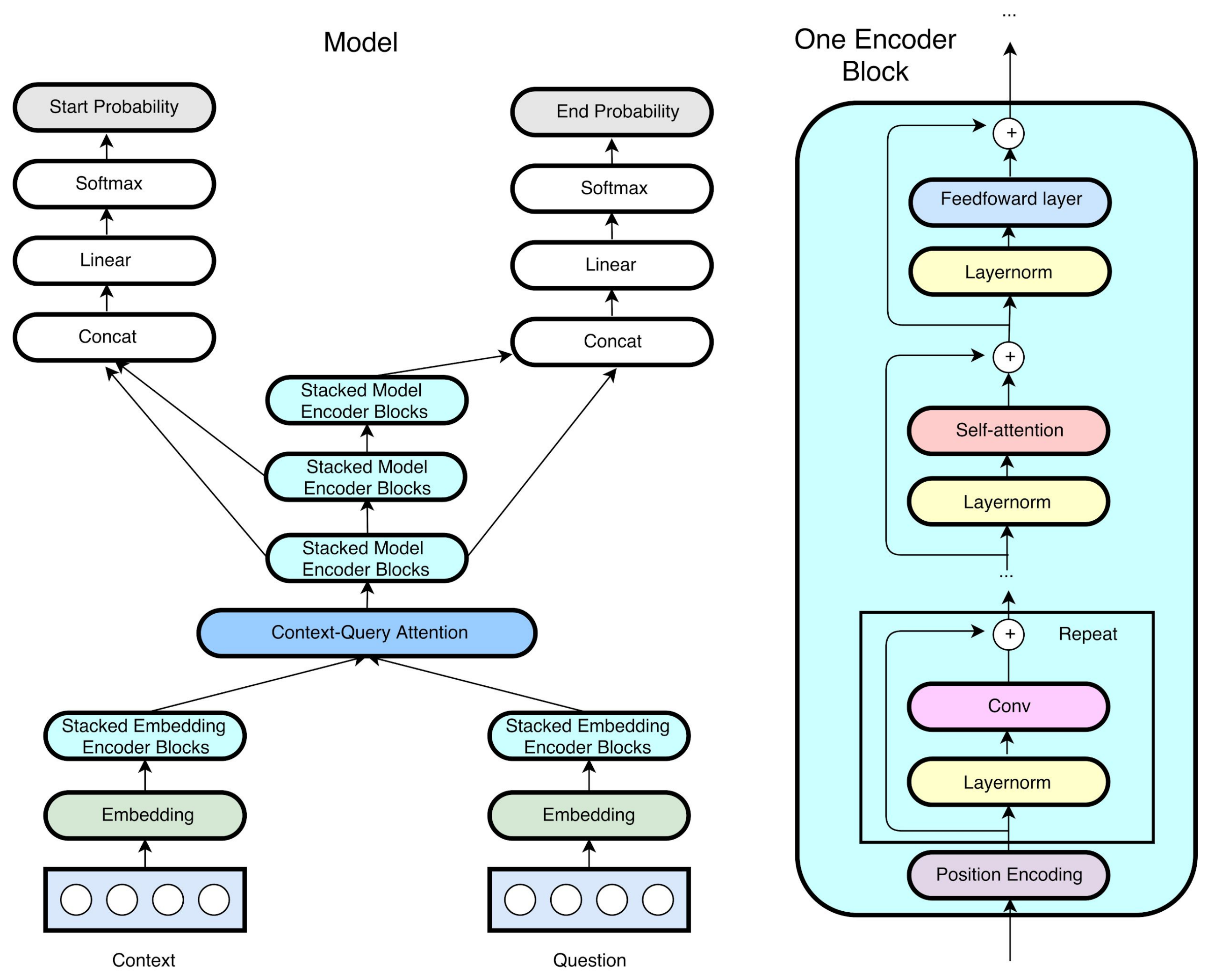
The task is, given a context paragraph and relevant question, to find the span of words within the context that answers the question, or output “no answer”. For example:

Context: *Frédéric François Chopin was a Polish and French composer and a virtuoso pianist of the Romantic era, who wrote primarily for the solo piano.*
Question: *The majority of Chopin's compositions were for what instrument?*
Answer: *solo piano*

Approach

I implemented QANet, an end-to-end neural model originally written for SQuAD 1.1. Its defining feature is the Encoder block, similar to the Transformer encoder, but which also uses ConvNets.

Context and question sequences are separately embedded and encoded, combined in a bidirectional attention layer, and encoded again before a softmax output layer that predicts the start and end of the answer.



The baseline was a version of BiDAF, a similar model to QANet but which uses LSTMs in the encoder layers instead of ConvNets and self-attention.

Results

The main evaluation metrics for SQuAD are EM and F1. EM is the percentage of model outputs that are exact matches for a reference answer, and F1 is the harmonic average of recall and precision.

The table to the right shows the results on the dev set of three models I trained, the original and improved BiDAF baseline models, and the QANet model.

Model	Dev EM	Dev F1
Baseline BiDAF (no char-embedding)	57.3	60.7
BiDAF (with char-embedding)	61.0	64.4
QANet	67.3	70.7

On the hidden test set, my QANet implementation obtained an EM score of 63.3 and an F1 score of 66.9, which was competitive on the class leaderboard.

Analysis

When the QANet model is evaluated on certain subsets of the dataset, the EM and F1 scores can vary. The table below shows scores when splitting the dataset by the first word of the question.

Question type	Count	Dev EM	Dev F1
“When...”	440	74.3	75.3
“Who...”	601	69.2	71.9
“Why...”	84	57.1	70.7
All questions	6078	67.3	70.7

Higher scores on “when” and “who” questions indicate skill at picking out years and named entities. A lower score on “why” questions may indicate some difficulty with understanding higher-level questions about a passage.

Conclusions

My implementation of QANet obtained reasonable results, and was competitive on the class leaderboard. This confirms the conclusions of the QANet paper of the model’s efficacy, and demonstrates that the QANet model is capable of good performance when adapted for SQuAD 2.0.

References

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.